Journal of Statistics Education

The Lure of Statistics in Data Mining

Lovleen Kumar Grover and Rajni Mehra
Guru Nanak Dev University, India, BBK DAV College for Women, India

Journal of Statistics Education Volume 16, Number 1 (2008),
www.amstat.org/publications/jse/v16n1/grover.html

Key Words: Censored data; Databases; Data dredging; Data fishing; Data mining; Exploratory data analysis; Knowledge discovery in data mining; Truncated data.

Abstract

The field of Data Mining like Statistics concerns itself with "learning from data" or "turning data into information". For statisticians the term "Data mining" has a pejorative meaning. Instead of finding useful patterns in large volumes of data as in the case of Statistics, data mining has the connotation of searching for data to fit preconceived ideas. Here we try to discuss the similarities and differences as well as the relationships between statisticians and data miners. This article is intended to bridge some of the gap between the people of these two communities.

1. Introduction

The two disciplines 'Statistics' and 'Data mining' are very similar. Statisticians and data miners commonly use many of the same techniques. More over statistical software vendors now include many of these techniques. Statistics developed as a discipline separate from Mathematics over the past century and a half to help scientists for making some sense of observations and to design experiments that yield the reproducible and accurate results associated with the scientific method. For almost all of this period, the issue was not too much data, but too little. Whereas, Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. Or, data mining is the application of Statistics in the form of exploratory data analysis and prediction models to reveal patterns and trends in very large datasets (see Berry-Linoff (1997)). Hence like Statistics, Data Mining is not only modeling and prediction but also a whole problem solving process.

The early development in the field of data mining was to facilitate a corporation for improving its marketing and sales. But nowadays, the data mining techniques and tools are equally applicable in other fields such as law enforcement, radio astronomy, medicines, and industrial process control etc. Many data mining algorithms have been developed but these algorithms were not invented keeping in mind any particular commercial application. The commercial data miner employs a grab bag of techniques borrowed from Statistics, Computer Science, and machine learning research. The choice of a particular combination of techniques to apply in a particular situation depends on the following:

(a) the nature of the data mining task,
(b) the nature of the available data, and
(c) the skills and preferences of the data miner.

As Statistics and Data Mining both deal with the process of learning from data so the main problem is to know how to get from data to information, from information to knowledge, from knowledge to decision, and from decision to action. Thus we can say, "We are drowning in information but starved for knowledge". The remedy to this problem is Data Mining and/or Statistics. Early statisticians were practical people who invented techniques to handle whatever problem was at hand. What is remarkable and a testament to the founders of modern statistics is that the earlier developed statistical techniques on tiny amount of data, have survived and still prove their utility. These techniques have proven their worth not only in the original domains but also in virtually all areas where data is collected from the fields. For example, the scientific study in the fields like agricultural experiments, psychological experiments, astronomical experiments, medicinal experiments and even in businesses etc. all requires the data in their respective fields to arrive at some logical decision. Stimulated by progress in computer technology and electronics data acquisition, recent decades have seen the growth of huge databases in many fields ranging from supermarket sales and banking, through astronomy, particle physics, chemistry, and medicines, to official and governmental statistics. These huge databases are viewed as a resource. It is certain that there is much valuable information in them, information that has not been tapped, and data mining is regarded as providing a set of tools by which that information may be extracted. The process of Data Mining is also known as KDD (Knowledge Discovery in Data mining). The main activities of Data Mining are: classification, estimation, prediction, affinity grouping or association rules, clustering; and description and visualization.

The advent of computing power has clearly simplified some aspects of analysis, although its bigger effect is probably the wealth of data produced. Our goal is no longer to extract every last iota of possible information from each rare datum. Our goal is instead to make sense of quantities of data so large that they are beyond the ability of our brains to comprehend in their raw format.

The present article starts with an introduction to what is probably the most important aspect of applied statistics-the skeptical attitude. It then discusses looking at data through a statistician's eye. This article ends with a brief discussion of some of the differences between data miners and statisticians and also differences in their attitude that are more a matter of degree than of substance.

2. Data Mining and Statistics

Many of the data mining techniques were invented by statisticians or have now been integrated into statistical software; they are extensions of standard statistics. Although, data miners and statisticians use similar techniques to solve similar problems, but the data mining approach differs from the standard statistical approach in several areas such as:

- Data miners assume that there is more than enough data and processing power.
- Data mining assumes dependency on time everywhere.

- It can be hard to design experiments in the business world without data mining.

These are the differences of approach, rather than opposites. As such, they shed some light on how the business problems addressed by data miners differ from the scientific problems that spurred the development of statistics. One major difference between business data and scientific data is that the latter is non-truncated /non-censored data and the former is truncated /censored data. Given a methodology or an algorithm to analyze data, it is often very hard to say whether it is "Statistics" or "Data Mining". It is not clear how one should put this label. Actually, in practice, while dealing with the real life problems in the industry, customers never ask, "Are you a data miner or a statistician?" In fact their main interest is in solving the problem in hand to their level of satisfaction and it is immaterial what label we use. As service men to the customers, we (as a data miner or as a statistician) need to try those statistical techniques or algorithms in our bag, which best suited to answer the queries of the customer.

By employing advanced analytical methods in Data Mining, businesses increase revenues, maximize operating efficiency, cut costs and improve customer satisfaction. Whereas Statistics make it possible to build predictive models or develop classifications that impact your bottom line. Without Statistics, there is no effective analysis. Without effective analysis, there is no business intelligence. Without business intelligence, how can you hope to assimilate gigabytes of data and consistently make decisions that will keep you ahead of your competition? With Statistics, you can transform your data in to knowledge about your business processes. Using Statistics in data mining can significantly impact all areas of your organization. Nowadays statistical software can improve your competitiveness from the shop floor to the sales floor to the executive floor. In today's business arena, it is a constant challenge to keep up with market trends and predict future outcomes. To increase market share and operate efficiently, you cannot afford not to use Statistics in Data Mining. If you are not mining your data for all it is worth, you are guilty of underuse of one of your company’s greatest assets. Although, there is subdiscipline within Statistics whose concern is description, a glance in any general statistics text will demonstrate that a central concern is how to make statements about a population when one has observed only a sample. However, data mining problem often have available the entire population of data, e.g. details of the entire workforce of a corporation, etc. In such cases, notion of statistically significance testing lose their meaning. On the other hand the central aim of data mining is discovery, it is not concern with those areas of statistics involving how best to collect the data in the first place so as to answer a specific question, such as experimental design and survey design. Data mining essentially assumes that the data have already been collected, and is concerned with how to discover its secrets. For further study about the similarities and differences between Statistics and data mining readers may consult Hand (1999a, 1999b).

3. Does Data Mining Mean Statistics or More Than Statistics?

Actually, data mining is a term synonymous with 'data dredging' or 'data fishing' and has been used to describe the process of trawling through data in the hope of identifying patterns. The data, which are not simply uniform, have differences that can be interpreted as patterns. The trouble is that many of these 'patterns' will simply be a product of random fluctuations, and will not represent any underlying structure. To statisticians, then, the term 'Data Mining' conveys the sense of naive hope vainly struggling against the cold realties of chance. To other researchers, however, the term is seen in a much more positive light. Superficially, of course, what we are describing here is nothing but exploratory data analysis, an activity that has been carried out since data were first analyzed and which achieved greater respectability. But there is a difference, and it is this difference that explains why statisticians have been slow to latch on to the opportunities. This difference is the sheer size of data sets now available. Statisticians have typically not concerned themselves with data sets containing many millions or even billions of records. Moreover, special storage and manipulation techniques required to handle them have been developed by entirely different intellectual communities from statisticians. It is probably no exaggeration to say that most statisticians are concerned with primary data analysis. On the other hand, data mining is entirely concerned with secondary data analysis. In fact we might define 'Data Mining' as the process of secondary analysis of large databases aimed at finding unsuspected relationships, which are of interest or value to the database owners. We see from this that 'Data Mining' is very much an inductive exercise, as opposed to the hypothetico-deductive approach often seen as the paradigm for how modern science progresses.

A common pattern is that a new idea will be launched by researchers in some other discipline, will attract considerable interest, and only then will statistician become involved. There is a real danger that Statistics and Statisticians will be perceived as a minor irrelevance, and as not playing the fundamental role in scientific and wider life that properly do. There is urgency for statistician to become involved with data mining exercises. Basically, classical statistics deals with numeric data. But nowadays, databases contain data of other kinds. Four obvious examples are image data, audio data, text data, and geographical data. The main issue of data mining consists of finding interesting patterns and structures in these databases. Of course, it is not possible simply to ask a computer to "search for interesting patterns" or to "see if there is any structure in the data". Before one can do this, one needs to define what one means by patterns or structure. And before one can do that one needs to decide what one means by "interesting". In general, of course, what is of interest will depend very much on the application domain. When searching for patterns or structures, a compromise needs to be made between the specific and the general. The essence of data mining is that one does not know precisely what sort of structure one is seeking, so a fairly general definition will be appropriate. On the other hand, too general a definition will produce too many candidate patterns. Since the pattern searches will throw up a large numbers of candidate patterns so there will be a high probability that spurious data configurations will be identified as patterns. Now, the problem is that how we deal with this situation? It is possible that a solution will only be found by stepping outside the conventional probabilistic statistical framework that is using scoring rules instead of probabilistic interpretations. The problem is similar to that of over fitting of statistical models, which is an issue that attracted renewed interest with the development of extremely flexible models such as neural networks.

In principle, a statistical expert system would embody a large base of intelligent understanding of the data analysis process, which it could apply automatically to a relatively small set of data. Whereas a data mining system, which embodies a small base of intelligent understanding, but which applies it to a large data set. In both cases the application is automatic, though in both cases interaction with the researcher is fundamental. In statistical expert system, the program drives the analysis following a statistical strategy because the user has insufficient statistical expertise to do so. Whereas in data mining application, the program drives the analysis because the user has insufficiently resources to manually examine billions of records and hundreds of thousands of potential patterns. For an elaborated view about data mining versus statistics, readers may consult Hand (1998). Given these similarities between the two enterprises, it is sensible to ask if there are lessons, which the data mining community might learn from the statistical expert system experience. The answer is certainly "Yes".

4. Should Data Mining be Included in the 'Statistics' Curriculum?

Statistics as taught in traditional curriculum may be described as being characterized by data that are small, clean, static and randomly sampled, and often collected to answer a specific question. None of these apply in the data-mining context. Because to a classical statistician, a data set with a few thousand observations may be large, but to a data miner, this is small.

It is possible to argue that, while there is great deal in common between data mining and statistics, the two have their own unique identities. We may also argue that the peculiarities of the problem they each tackle and the nature and constraints of the methods they utilize could lead to a fruitful synergy. In fact, there are deep theoretical issues arising from data mining problems, which would benefit from a statistical perspective and understanding. It appears that, data mining may be put within the context of greater statistics that can be defined, at least loosely, as "every thing related to learning from data". Greater statistics tends to be inclusive, eclectic with respect to methodology, closely related to other disciplines, and practiced by many outside of traditional professional statistics and outside of academia.

The majority of data miners appear to have relatively little formal statistical expertise. Hence, they sometimes make errors which trained statisticians would avoid as obvious. This implies that data miners must take on board statistical insights regarding the potentials for spurious associations and issues of substance versus statistical significance, leading to a requirement for training data miners in statistics or statistics graduates in data mining. The approach needs to be practical and example-based, and some re-focusing of traditional statistics curriculum is desirable emphasizing changes in data collection and analysis, which have emerged in the past fifteen or so years. The last decade or so have seen hundred of computer software manufacturer jumping onto the data mining bandwagon. Major statistical software packages such as SAS, S-PLUS, SPSS, and STATISTICA, etc. are being marketed as data mining tools rather than statistical tools.

Computer scientists have beaten the statisticians in offering data mining courses, since the advent of substantial improvement in computing power, in the 1990s. However, most if not all of these offerings have concentrated on the implementation of efficient algorithms from a machine learning point of view. Although, several service-oriented and problem-driven statistical research methods courses are being offered at tertiary level, often via statistics units or consulting centers, courses on statistics oriented data mining has not been widely available on the menu. These courses should provide a broad statistical perspective of data mining at undergraduate level, and are aimed at students majoring in statistics and at those majoring in fields such as computer science, database management and business studies. Such courses must provide a broad coverage of techniques often categorized as supervised or unsupervised, and provide descriptive or predictive modeling introduction. A typical prescription may take the form, "...an introduction to data mining applications including data preparation and data warehousing; query, association, market basket and rule induction methods; prediction using regression, decision trees and neural nets; clustering using hierarchical methods and self-organizing maps; classification using trees and neural nets; a visual approach with real examples and case studies; use of leading data mining software tools; ... ." An advanced course in statistical data mining, say at postgraduate level, may include statistical underpinning of commonly used techniques plus a wide range of new developments such as genetic algorithms, text-mining, bagging, bumping and boosting algorithms, and Bayesian belief networks.

Actually the main question is, if a student in statistics curriculum wants to take up a career in applied statistics in industry, what should we teach him or her? How best can we equip him/her to be most effective in the future role? Along with regular statistics courses there should be some courses related to various data mining algorithms and a study about the use of software, which include the implementation of these algorithms. We should expose students of statistics to various databases, various types of data sets from different domains. The students of statistics should also be aware about the challenges of storing, accessing and manipulating large volumes of data with effective visualization and presentation.

There is no doubt that there is mutual ignorance between statisticians and data miners. One part of the reason for this mutual ignorance lies in conservativeness in statistics versus a risk-taking attitude in computing. There is now wide acceptance that progress in data mining will demand a merging of the insights of computing specialists with those of statisticians. It is something of an indictment of the statistical profession that few statisticians have become involved in a deep way with data mining. Statisticians have a lot to teach data miners, while data miners have many fascinating and new problems, which statisticians have not even begun to look at. There is the opportunity for an immensely rewarding synergy between statisticians and data miners. However, most data miners tend to be ignorant of statistics and client's domain; statisticians tend to be ignorant of data mining and client's domain; and clients tend to be ignorant of data mining and statistics. Unfortunately, they also tend to be inhibited by myopic points of view; computer scientists focus upon database manipulations and processing algorithms; statisticians focus upon identifying and handling uncertainties: and clients focus upon integrating knowledge into the knowledge domain. Moreover, most data miners and statisticians continue to sarcastically criticize each other. This is detrimental to both disciplines. Unfortunately, the anti-statistical attitude will keep data mining from reaching its actual potential – data mining can learn from statistics. Data mining and statistics will inevitably grow toward each other in the near future because data mining will not become knowledge discovery without statistical thinking, statistics will not be able to succeed on massive and complex datasets without data mining approaches. Remember that knowledge discovery rests on the three balanced legs of computer science, statistics and client knowledge; it will not stand either on one leg or on two legs, or even on three unbalanced legs. Hence, successful knowledge discovery needs a substantial collaboration from these three communities. All parties should widen their focus until true collaboration and the mining for gold becomes reality. A maturity challenge is for data miners, statisticians and clients to recognize their dependence on each other and for all of them to widen their focus until true collaboration becomes reality. The critical challenge for us all is to view the challenges as opportunities for our joint success. All parties should widen their focus until true collaboration and the mining for gold becomes reality. It would be nice to think that researchers from these disciplines would come together to pool their distinct perspectives and approaches, to tackle the really important problems which we are facing in this modern data-rich world. There is no doubt that data mining is 'statistically intellectual'. For further discussion on this particular issue readers may consult Ganesh (2002) and Kuonen (2004). Statistical exploration of data mining has been active in recent years by means of research articles appearing in statistical as well as non-statistical journals. The statistical training brings to the table for any data analysis/mining problem has the following key contributions:

- The notion of probabilistic models
- Idea of measurement errors
- Idea of statistical significance
- Difference due to root cause versus difference due to pure chance

In terms of problem at hand, these have direct impacts like generalizabilty of the results, quality of recommendations for a possible process improvement etc. Such things are typically not dealt with a great depth in any non-statistics course. It is important to realize that not all dataset are massive. Many times, it will fit nicely in their local desktop or laptop's memory. In view of this, it is all the more important for data analysts to focus more on mastering the science and art of modeling and analysis. So courses on statistics oriented data mining should be provided in the 'statistics' curriculum. However the debatable issue in the near future is "Whether statistics as a field should embrace data mining as subdicipline or leave it to the computer scientists?

Finally, turning to the question "Should data mining be included in the statistics curriculum?", the answer inevitably should be an enthusiastic “Yes.” After all, Data mining essentially is statistical data analysis. As the experience tells us thatStatistics as a discipline has a poor record for timely recognition of important ideas, so definitely there is a chance to improve the reputation of this discipline. It would be a great loss, for the reputation of Statistics as a discipline as well as for individual Statisticians, if these opportunities were not grasped.

5. Conclusions

The statisticians and the data miners solve similar problems. However, because historical differences and differences in the nature of the problems, there are some differences in approaches. There is clear potential, opportunity, and indeed even excitement in data mining for making discoveries in large data set. A lively debate on the issue "Difference between statistics and data mining" concluded that it is immaterial what you call it either data mining or statistics. Since 'computation' plays a major role in the process of data mining so computer scientists have significant claim over the ownership of data mining. Nevertheless, data mining techniques, in general, have a statistical base; and statisticians are beginning to show a significant interest in the area, including offering tertiary courses in statistical data mining. Further general reading on data mining and statistics may be found in the references such as Berry and Linoff (1997, 2000), Chatfield (1997), Friedman (1998), Hand (1999a, 1999b), Hastie et al. (2001).

Acknowledgement

Many thoughts and statements in this article are courtesy of various authors of texts, papers from journal and conference proceedings cited below. We are very much grateful to the unknown referees for their valuable suggestions, which substantially improved the quality of this paper. We are also thankful to the editor of *JSE* for his keen interest in this paper.

References

Berry, J.A.M., and Linoff, G. (1997), *Data mining techniques-for marketing, sales and customer support*, New York: Wiley.

Berry, J.A.M., and Linoff, G. (2000), *Mastering data mining -the art and science of customer relationship management*, New York: Wiley.

Chatfield, C. (1997), “Data mining”, *Royal statistical society news*, 25, 1-2.

Friedman, J.H (1998), “Data mining and statistics-what's the connection”, 29th Symposium on the interface.

Ganesh, S. (2002), “Data mining: Should it be included in the statistics curriculum?” The 6th international conference on teaching statistics (ICOTS 6), Cape Town, South Africa.

Hand, D.J. (1998), “Data mining-statistics and more?”, *The American Statistician*, 52 112-118.

Hand, D.J. (1999a):, "Data mining-new challenges for statisticians", Proceedings of the ASC (Association for survey computing) International conference, 21-26.

Hand, D.J. (1999b), "Statistics and Data mining-intersecting disciplines", *SIGKDD Explorations*, 1, 16-19.

Hastie, T.; Tibshirani, R. and Friedman, J.H. (2001), *Elements of statistical learning-data mining inference and prediction*, New York: Springer Verlag.

Kuonen, D. (2004), “Data mining and Statistics: What is the connection?”, *The Data Administrative Newsletter*, Switzerland.

Lovleen Kumar Grover
Department of Mathematics
Guru Nanak Dev University
Amritsar 143 005, Punjab
India
lovleen_2@yahoo.co.in

Rajni Mehra
Department of Computer Science
BBK DAV College for Women
Amritsar 143 005, Punjab
India
rajni_mehra7@yahoo.co.in

Volume 16 (2008) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications